Small Area Estimation: Methods and Applications

J. N. K. Rao

Carleton University, Ottawa, Canada

Paper presented at the Seminar “Applications of Small Area Estimation Techniques in the Social Sciences”, October 3 – 5, 2012, Iberoamerican University, Mexico City

Introduction

- Censuses have limited scope. Sample surveys can provide reliable current statistics for large areas or subpopulations (domains).

- Growing demand for reliable small area statistics but sample sizes are too small to provide direct (or area specific) estimators with acceptable accuracy.

- Examples of small areas: county, municipality, three digit occupation counts within a province, health regions, even a state by age-sex-race groups.

- Domain or subpopulation is called a small area if the domain-specific sample size is small.

- It is necessary to borrow strength from related areas through linking models based on auxiliary data such as census data and administrative records. This leads to indirect estimators.

- Major application (SAIPE) Estimation of counts of school age children under poverty in USA at the county and

school district level. More than 15 billion dollars of federal funds are allocated on the basis of model-based indirect estimators constructed from the Current Population Survey (CPS) data and administrative records and census data.

Design issues

- Can we minimize the use of indirect estimators by taking preventive measures at the design and/or estimation stage?

- At the design stage, possible actions: (a) Replace clustering: use list frames (b) Use many strata (c)

Compromise sample allocation: Canadian LFS used a two-step allocation: 42000 optimized at the provincial level and 17000 at UI level. Consequence: Maximum coefficient of variation (CV) reduced from 18% to 9% at UI level with slight increase in CV at the provincial level. (d) Optimal allocation: Minimize total sample size subject to desired tolerances on the CVs of direct (or indirect) estimators of areas and aggregate of areas (Choudhry, Rao and Hidiroglou 2012) (e) Integration of surveys, multiple frames, rolling samples.

- Estimation stage: use efficient direct estimators.

- "The client will always require more than is specified at the design stage" (Fuller 1999). So we cannot avoid unplanned small domains.

Models for small area estimation

- **Parameters of interest:** Area means, totals, quantiles, proportions. Complex measures: Poverty indicators (for example, rate, gap and severity used by the World Bank).

- **Area-level Fay-Herriot model:** Direct estimates for areas and area-level auxiliary data available. Consider simple random sampling within areas ($i = 1, \ldots, m$) and

sample mean $\bar{y}_i$ as direct estimators of the population mean $\bar{\bar{Y}}_i$. In practice, direct estimators could be based on calibration etc. Some areas of interest may not be sampled at all. For example, in the SAIPE application only 1000 of the 3000 counties are sampled in the CPS.

- **Sampling model:** $\bar{y}_i = \bar{\bar{Y}}_i + e_i$,

$$e_i \sim_{ind} N(0, \psi_i), \psi_i \text{ known}$$

- **Linking model:** $\bar{\bar{Y}}_i = z_i'\beta + v_i \qquad v_i \sim_{iid} N(0, \sigma_v^2)$

Parameters of interest: $\bar{\bar{Y}}_i$

z_i = Vector of area-level covariates obtained from census and administrative sources

- **Combined model:** $\bar{y}_i = z_i'\beta + v_i + e_i,\ i = 1,...,m$

Special case of a linear mixed model

Optimal estimation of $\bar{Y}_i$

- Best (Bayes) estimator under the assumed model is the conditional expectation of $\bar{Y}_i$ given the survey estimator $\bar{y}_i$ and model parameters β and σ_v^2:

$$\hat{\bar{Y}}_i^B = \gamma_i \bar{y}_i + (1-\gamma_i) z_i'\beta,\ \gamma_i = \sigma_v^2 / (\sigma_v^2 + \psi_i)$$

- Best estimator is a weighted combination of direct estimator $\bar{y}_i$ and regression synthetic estimator $z_i'\beta$. More weight is given to direct estimator when the sampling variance is small relative to total variance and more weight to the synthetic estimator when sampling variance is large or model variance is small.

- In practice, we need to estimate model parameters using MM (moment methods: Fay-Herriot, 1979), ML or REML. Resulting estimator is called **Empirical Best or Empirical Bayes (EB)**:

$$\hat{\bar{Y}}_i^{EB} = \hat{\gamma}_i \bar{y}_i + (1-\hat{\gamma}_i) z_i' \hat{\beta}$$

- MM does not require normality assumption and the resulting EB estimator is also EBLUP (empirical best linear unbiased prediction) estimator.
- Customary weighted least squares estimator of β may not be the best from a prediction point of view unless the mean model $z_i'\beta$ is correctly specified. Jiang, Nguyen and J. S. Rao (2011) proposed alternative estimator of β that makes the EB estimator perform well under misspecification of the mean model.

- Estimation of σ_v^2 often leads to difficulties due to negative estimates which are truncated to zero. This is an area of active research and one recent method, called adjusted ML method under normality, gives strictly positive estimates (Li and Lahiri, 2010).

- Preliminary test estimation (Datta et al. 2010): Test for the hypothesis $H_0 : \sigma_v^2 = 0$ using significance level $\alpha = 0.2$. If H_0 is not rejected, use synthetic estimator under the model without the random effects, otherwise use the EB estimator.

- Assumption of known sampling variances ψ_i is also a problem.

- For non-sampled areas with known z_i, regression synthetic estimator $z_i'\hat{\beta}$ is used.

- A more realistic linking model is to replace $\bar{Y}_i$ by $h(\bar{Y}_i)$ for some suitable function $h(.)$. For example, use a logistic transformation if the mean is a proportion. In this case, sampling model and linking model are mismatched and hence cannot be combined into a linear mixed model. More sophisticated methods are needed (You and Rao 2002). Mohadjer et

al. (2012) applied mismatched models to estimate the county proportions of adults at the lowest literacy level using hierarchical Bayes (HB) methods.

- Sampling model is modified to match the linking model by taking it as $h(\bar{y}_i) = h(\bar{Y}_i) + v_i + \tilde{e}_i$, but the mean of induced sampling error $\tilde{e}_i$ may not be zero. Find the EB estimator of $h(\bar{Y}_i)$ and then back transform to get the estimator of $\bar{Y}_i$ which is not EB.

- Benchmarking: (a) Adjust the EBLUP estimators to make the estimators agree with a reliable direct estimator at an aggregate level. For example, use ratio

benchmarking. (b) Self-benchmarking: Use a different estimator of β to ensure benchmarking (You, Rao and Hidiroglou 2012) or augment the model using $w_i \psi_i$ as the additional covariate where w_i is the weight used for aggregation (Wang et al. 2008).

- **Mean squared prediction error**

$\text{MSPE}(\hat{\bar{Y}}_i^{EB}) = E(\hat{\bar{Y}}_i^{EB} - \bar{Y}_i)^2$ is used as a measure of precision of the EB estimator. If the number of areas m is moderately large then

$$\text{MSPE}(\hat{\bar{Y}}_i^{EB}) \approx g_{1i}(\sigma_v^2) + g_{2i}(\sigma_v^2) + g_{3i}(\sigma_v^2)$$

- Leading term $g_{1i}(\sigma_v^2) = \gamma_i \psi_i$ shows large reduction in MSPE can be achieved relative to the direct estimator $\bar{y}_i$ if γ_i is small. Second term is due to the estimation of β and the last term is due to estimation of σ_v^2 and they are of lower order.

- MSPE estimator is obtained by replacing σ_v^2 by its estimator and multiplying the last term by the factor 2.

- Alternative methods use resampling including jackknife and bootstrap which are more computer intensive but more widely applicable. MSPE estimation is an area of active research.

- For the preliminary test estimator, use the MSPE estimator under the model without random effects when H_0 is not rejected, otherwise use the usual MSPE estimator (Datta et al. 2010).

Model selection and checking:

Variable selection: Fence method (Jiang et al. 2008), Conditional Akaike Information Criterion (AIC) for predictive performance (Han 2011)

Model checking: Residual analysis (weighted Q-Q plots, influential diagnostics: Rao 2003, Chapters 6 and 7)

Applications of area level model

(1) Estimation of per capita income (PCI) for small places in the United States (Fay and Herriot 1979)

- Direct estimates $\bar{y}_i$ of PCI from the sample census are not reliable for small places. Uses $h(\bar{y}_i) = \log(\bar{y}_i)$ and log (PCI for the county) and log (value of housing for the place) as predictor variables.
- A method of model checking from the sample data proposed. Also external evaluation by comparing to true values from a special census of a sample of small places.

(2) SAIPE for counties

- Direct county estimates of total poor obtained from the Current Population Survey and log (total poor) taken as the response variable in the model. Prediction variables include log of the following: food stamps, poor from tax forms, number of exemptions, last census poor.
- Extensive internal model checking ignoring the random area effect and using fixed effects model. External evaluations using census estimates (Chapter 7, Rao 2003).

Borrowing strength over time: Time series and cross-sectional model (Rao and Yu 1994)

- Sampling model: $\bar{y}_{it} = \bar{Y}_{it} + e_{it}$

  Linking model: $\bar{Y}_{it} = z'_{it}\beta + v_i + u_{it},\ v_i \sim N(0, \sigma_v^2)$,

$u_{it} = \rho u_{i,t-1} + \varepsilon_{it},\ \varepsilon_{it} \sim N(0, \sigma^2)$, AR (1): $|\rho| < 1$

- Random walk: $\rho = 1$

Application: Estimation of median income of four-person families for the states in USA. Direct estimates from CPS and predictor variable is the adjusted census income. CV of direct estimates exceeded 6% for 38 states compared to 0% for EBLUP estimates.

Unit level models: Nested error regression model

- Suppose we select simple random samples of sizes n_i from the areas i with sizes N_i and observe unit responses $\{(y_{i1},...,y_{in_i});i=1,...,m\}$ and associated covariates x_{ij} with known area means $\bar{X}_i$. Objective is to estimate the small area means $\bar{Y}_i$.

Model

$$y_{ij}=x_{ij}'\beta+v_i+e_{ij},\ j=1,...,n_i;i=1,...,m$$

$$v_i \sim_{iid} N(0,\sigma_v^2), e_{ij} \sim_{iid} N(0,\sigma_e^2),\ \ v_i \perp e_{ij}$$

- Area mean $\bar{Y}_i \approx \bar{X}_i'\beta + v_i$ if n_i is small relative to N_i where the area population means $\bar{X}_i$ known from census or administrative records.

- EB estimator of the mean $\bar{Y}_i$ is a weighted average of the sample regression estimator $\bar{y}_i + \hat{\beta}(\bar{X}_i - \bar{x}_i)$ and the regression synthetic estimator $\bar{X}_i'\hat{\beta}$ with weights $\hat{\gamma}_i = \hat{\sigma}_v^2 / (\hat{\sigma}_v^2 + \hat{\sigma}_e^2 / n_i)$ and $1 - \hat{\gamma}_i$ respectively. EB estimator can be written as

$$\hat{\bar{Y}}_i^{EB} = \bar{X}_i\hat{\beta} + \hat{v}_i = \sum_{j \in U_i} \hat{y}_{ij}$$

$$\hat{y}_{ij} = x_{ij}'\hat{\beta} + \hat{v}_i$$

- If the sampling fraction n_i / N_i is not negligible, then EB is modified to

$$\hat{\bar{Y}}_i^{EB} = N_i^{-1}\{\sum_{j\in s_i} y_{ij} + \sum_{j\in U_i - s_i} \hat{y}_{ij}\}$$

EB estimator is also EBLUP without normality assumption using MM to estimate model parameters.

- Estimation of MSPE is similar to the area level case.

Design consistency and benchmarking

- Pseudo-EBLUP estimator obtained from an aggregated model based on design weights ensures design consistency

and self-benchmarking to a reliable direct estimator of a larger area covering small areas (You and Rao 2002).

- Informative sampling: Population model does not hold for the sample. Current research provides tools to handle this case (Pfeffermann and Sverchkov 2007). Augmenting the nested error model by including survey weights as additional covariates and then doing EB is a simple method and seems to work well (Verret, Hidiroglou and Rao 2012). But it assumes that the sum of the population weights for each area is known.

What can go wrong?

(1) Outliers in random effects v_i and /or unit errors e_{ij}: Use robust EBLUP assuming mean zero random effects and unit errors (Sinha and Rao 2009).

(2) Relax the mean assumption by replacing mean function $x_{ij}'\beta$ in the model by some smooth function and approximate it by a P-spline. Resulting model has a linear mixed model form so use EBLUP (Opsomer et al. 2008). Robust EBLUP version to handle outliers: Rao et al. (2010).

(3) What if the specified model is wrong? Use a model assisted approach by treating the model as a working model and then doing design-based bias correction (Lehtonen and Veijanen 1999). Under simple random sampling within areas, estimator of mean $\bar{Y}_i$ is given by

$$\hat{\bar{Y}}_i^{EB,c} = N_i^{-1}\{\sum_{j\in U_i} \hat{y}_{ij} + (N_i / n_i)\sum_{j\in s_i} (y_{ij} - \hat{y}_{ij})\}$$

- Note that if n_i is small the bias-corrected estimator will have large CV because the bias correction is a direct estimator based only on the sample s_i in area i

EB estimation of small area poverty indicators

- E_{ij} is a welfare measure for individual j in area i and z is the poverty line.
- World Bank (WB) family of poverty indicators:

$$F_{\alpha ij} = \{(z - E_{ij})/z\}^{\alpha} I(E_{ij} < z) \Rightarrow F_{\alpha i} = N_i^{-1} \sum_{j \in U_i} F_{\alpha ij}$$

$\alpha = 0$: Poverty incidence, $\alpha = 1$: poverty gap

- Also called FGT poverty measures (Foster et al. 1984)

- Transform E_{ij} to $y_{ij} = \log(E_{ij})$ and express $F_{\alpha i}$ as a function of the y_{ij}, say $h_\alpha(y_i)$.
- EB estimator of $F_{\alpha i}$ = conditional expectation of $h_\alpha(y_i)$ with respect to the estimated predictive distribution of non-sampled y_{ir} given the sampled y_{is}.

Implementation (Monte Carlo approximation)

- Generate L non-sampled $y_{ir}^{(l)}, l = 1,..., L$ from the estimated predictive distribution.
- Attach the sampled elements to form simulated census vectors $y_i^{(l)}, l = 1,..., L$.

- Calculate the desired poverty measure with each population vector: $F_{\alpha i}^{(l)} = h_\alpha(y_i^{(l)}), l = 1,..., L.$
- Take the average over the L simulated censuses as an approximation to the EB estimator:
- Under the nested error model on y_{ij} and normality, we can generate values from the estimated predictive distribution using only univariate normal distributions (Molina and Rao 2010).

$$F_{\alpha i}^{EB} \approx L^{-1} \sum_{l=1}^{L} F_{\alpha i}^{(l)}$$

- WB uses another simulated census method but the proposed EB method can be considerably more efficient than the WB method for sampled areas.

Current Research

- Relax normality assumption by assuming a family of skew normal distributions for the random effects and the errors in the nested error model.
- WB is also studying EB method using a mixture of normal distributions approach.

Hierarchical Bayes (HB) approach

- HB method provides “exact” inferences including point estimator, posterior variance and credible intervals, assuming that the model parameters are random with “non-informative” priors (Chapter 10, Rao 2003). Predictive distribution of the parameter of interest is obtained for this purpose.
- Molina, Nandram and Rao (2012) applied the HB approach to poverty indicators. They proposed a method of drawing samples from the predictive distribution of non-sampled given the sampled values without resorting to the frequently used MCMC

methods. Simulation study demonstrated that the HB method has good frequentist properties.

Applications of HB

- HB estimation of adult literacy for states and counties using mismatched sampling and linking models (Mohadjer, Rao et al. 2012)

Alternative Methods

- M-quantile estimators: random small area effects are not directly incorporated into the model (Chambers and

Tzavidis 2006). Estimators are essentially synthetic if the area sampling fraction is small. Bias reduction methods are also applied to M-quantile estimators (Tzavidis et al. 2010).

Recommendations

(1) Preventive measures at the design stage may reduce the need for indirect estimators significantly.

(2) Good auxiliary information related to the variables of interest plays a vital role in model-based estimation. Expanded access to auxiliary data, such as census and

administrative data, through coordination and cooperation among federal agencies is needed.

(3) Model selection and checking plays an important role. External evaluations are also desirable whenever possible.

(4) Area-level models have wider scope because area-level data are more readily available. But assumption of known sampling variance is restrictive.

(5) HB approach is powerful and can handle complex modeling, but caution should be exercised in the choice of priors on model parameters. Practical issues in implementing HB paradigm should be addressed (Rao 2003, Section 10.2.4)

(6) Model-based estimates of area totals and means not suitable if the objective is to identify areas with extreme population values or to identify areas that fall below or above some pre-specified level (Rao 2003, Section 9.6).

(7) Suitable benchmarking is desirable.

(8) Model-based estimates should be distinguished clearly from direct estimates. Errors in small area estimates may be more transparent to users than errors in large area estimates.

(9) Proper criterion for assessing quality of model-based estimates is whether they are sufficiently accurate for the

intended uses. Even if they are better than direct estimates, they may not be sufficiently accurate to be acceptable.

(10) Overall program should be developed that covers issues related to sample design and data development, organization and dissemination, in addition to those pertaining to methods of estimation for small areas.

References

Chambers, R. and Tzavidis, N. (2006). M-quantile models for small area estimation. *Biometrika*, 73, 597-604.

Choudhry, G. H., Rao, J. N. K. and Hidiroglou, M. A. (2012)). On sample allocation for efficient domain estimation. *Survey Methodology*, 38, 23-29.

Datta, G. S., Hall, P. And Mandal, A. (2011). Model selection for the presence of small-area effets and applications to area-level data. *Journal of the American Statistical Association*, 106, 362-374.

Fay, R.E. and Herriot, R. A. (1979). Estimates of income for small places: an application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74, 269-277.

Foster, J., Greer, J. and Thorbecke, E. (1984). A class of decomposable poverty measures. *Econometrica*, 52, 761-766.

Fuller, W. A. (1999). Environmental surveys over time. *Journal of the Agricultural, Biological and Environmental Statistics*, 4, 331-345.

Jiang, J., Rao, J. S., Gu, Z and Nguyen, T. (2008). Fence methods for mixed model selection. *Annals of Statistics*, 1669-1692.

Jiang, J., Nguyen, T. and Rao, J. S. (2011). Best predictive small area estimation. *Journal of the American Statistical Association*, 106, 732-745.

Lehtonen, R., Sarndal, C. E. and Veijanen, A. (2003). The effect of model choice in estimation for domains. *Survey Methodology*, 29, 33-44.

Li, H. and Lahiri, P. (2010). An adjusted maximum likelihood method for solving small area estimation problems. *Journal of Multivariate Analysis*, 101, 882-892.

Mohadjer, L, Rao, J. N. K., Liu, B., Krenzke, T. and Van de Kerckhove, W. (2012). Hierarchical Bayes small area estimates of adult literacy using unmatched sampling and linking models. *Journal of the Indian Society of Agricultural Statistics*, 66, 55-63.

Molina, I. and Rao, J. N. K. (2010). Small area estimation of poverty indicators. *Canadian Journal of Statistics*, 38, 369-385.

Molina, I., Nandram, B. and Rao, J. N. K. (2012). Hierarchical Bayes small area estimation of general parameters with application to poverty indicators. Technical Report.

Opsomer, J.D., Claeskens, G., Ranalli, M. G. Kauemann, G. and Breidt, F. J. (2008). Non-parametric small area estimation using penalized spline regression. *Journal of the Royal Statistical Society*, series B, 70, 265-286.

Pfeffermann, D. and Sverchkov, M. (2007). Small-area estimation under informative probability sampling. *Journal of the American Statistical Association*, 102, 1427-1439.

Rao, J. N. K. and Yu, M. (1994). Small area estimation by combining time series and cross-sectional data. *Canadian Journal of Statistics*, 22, 511-528.

Rao, J. N. K. (2003). *Small Area Estimation*. Wiley, Hoboken, New Jersey.

Rao, J. N. K., Sinha, S. K. and Roknossadati, M. (2009). Robust small area estimation under penalized spline mixed models. In Proceedings of the Survey Research Section, American Statistical Association, pp. 145-153.

Sinha, S. K. and Rao, J. N. K. (2009). Robust small area estimation. *Canadian Journal of Statistics*, 37, 381-399.

Tzavidis, N., Marchetti, S. and Chambers, R. (2010). Robust estimation of small-area means and qunatiles. *Australian and New Zealand Journal of Statistics*, 52, 167-186.

You, Y. and Rao, J. N. K. (2002). A pseudo-empirical best linear unbiased prediction approach to small area estimation using survey weights. *Canadian Journal of Statistics*, 30, 431-439.

You, Y., Rao, J. N. K. and Hidiroglou, M. (2012). On the performance of self benchmarked small area estimators under the Fay-Herriot area level model. *Survey Methodology*, 38 (in press).

Verret, F., Hidiroglou, M. A. and Rao, J. N. K. (2012). Model-based small area estimation under informative sampling. Technical Report.

Wang, J., Fuller, W. A. and Qu, Y. (2008). Small area estimation under a restriction. *Survey Methodolgy*, 34, 29-36.

You, Y. and Rao, J. N. K. (2002). A pseudo-empirical best linear unbiased prediction approach to small area estimation using survey weights. *Canadian Journal of Statistics*, 30, 431-439.

You, Y., Rao, J. N. K. and Hidiroglou, M. (2012). On the performance of self benchmarked small area estimators under the Fay-Herriot area level model. *Survey Methodology*, 38 (in press).